대한민국특허청
KOREAN INTELLECTUAL PROPERTY OFFICE

별첨 사본은 아래 출원의 원본과 동일함을 증명함.

This is to certify that the following application annexed hereto is a true copy from the records of the Korean Intellectual Property Office.

출 원 번 호 : 10-2003-0056947
Application Number

출 원 년 월 일 : 2003년 08월 18일
Date of Application AUG 18, 2003

출 원 인 : 학교법인 한국정보통신학원
Applicant(s) INFORMATION AND COMMUNICATIONS UNIVERSITY EDUCA

2004 년 01 월 13 일

특 허 청
COMMISSIONER

【서지사항】

【서류명】	특허출원서
【권리구분】	특허
【수신처】	특허청장
【제출일자】	2003.08.18
【발명의 명칭】	도메인 조합 정보를 이용한 단백질간의 상호 작용 예측 방법 및 예측 시스템
【발명의 영문명칭】	SYSTEM AND METHOD FOR PREDICTING THE INTERACTION BETWEEN PROTEINS BASED ON DOMAION COMBINATION
【출원인】	
【명칭】	학교법인 한국정보통신학원
【출원인코드】	2-1999-038195-0
【대리인】	
【성명】	정태영
【대리인코드】	9-2001-000339-7
【발명자】	
【성명의 국문표기】	한동수
【성명의 영문표기】	HAN,Dong-Soo
【주민등록번호】	621201-1481216
【우편번호】	305-762
【주소】	대전광역시 유성구 전민동 엑스포아파트 401-1602
【국적】	KR
【발명자】	
【성명의 국문표기】	김홍숙
【성명의 영문표기】	KIM,Hong-Soog
【주민등록번호】	690126-1011634
【우편번호】	305-728
【주소】	대전광역시 유성구 전민동 세종아파트 101동405호
【국적】	KR
【발명자】	
【성명의 국문표기】	서정민
【성명의 영문표기】	SEO,Jung-Min
【주민등록번호】	710104-2691413

【우편번호】	305-150
【주소】	대전광역시 유성구 반석동 418번지 101호
【국적】	KR
【발명자】	
【성명의 국문표기】	장우혁
【성명의 영문표기】	JANG,Woo-Hyuk
【주민등록번호】	800707-1690514
【우편번호】	702-250
【주소】	대구광역시 북구 동천동 915 칠곡3차 화성아파트 104-1309
【국적】	KR
【심사청구】	청구
【조기공개】	신청
【취지】	특허법 제42조의 규정에 의한 출원, 특허법 제60조의 규정에 의한 심사청구 , 특허법 제64조의 규정에 의한 출원공개를 신청합니다. 대리인 정태영 (인)
【수수료】	
【기본출원료】	20 면 29,000 원
【가산출원료】	12 면 12,000 원
【우선권주장료】	0 건 0 원
【심사청구료】	7 항 333,000 원
【합계】	374,000 원
【감면사유】	학교
【감면후 수수료】	187,000 원
【첨부서류】	1. 위임장[추후제출]_1통 2.고등교육법 제2조에의한 학교임을 증명하는 서류_1통

【요약서】

【요약】

본 발명은 두 개의 단백질 간에 상호 작용이 있을 가능성을 통계적으로 예측하는 방법 및 시스템에 관한 것으로, 특히 단백질간 상호 작용이 있는 단백질쌍 집단과 상호 작용이 없는 단백질쌍 집단간의 차이를 도메인 조합(Domain Combination)의 관점에서 분석한 정보에 기반하여 미지의 단백질쌍의 상호 작용할 가능성을 예측하기 위한 것이다.

이를 위하여 본 발명에 의한 단백질간 상호 작용의 예측 방법은 상호 작용 단백질쌍 모집단과 비상호 작용 단백질쌍 모집단의 각 집단으로부터 각각 선정한 특정 도메인 조합의 출현 확률 정보를 추출하고 저장하는 단계와, 상기 저장된 도메인 조합의 출현 확률 정보를 이용하여 임의의 두 단백질이 상호 작용할 확률식을 결정하는 단계, 및 상기 결정된 확률식으로부터 임의의 두 단백질이 상호 작용할 확률을 구하는 단계를 포함하는 것을 특징으로 한다.

본 발명에 의할 때, 도메인 조합 정보를 이용하여 단백질간 상호 작용 가능성을 예측함으로써 단일 도메인 쌍(Single Domain Pair)을 이용하는 종래의 예측방법 및 시스템과 비교하여 매우 향상되고 개선된 예측력을 기대할 수 있다.

【대표도】

도 1

【색인어】

도메인(Domain), 도메인 조합(Domain Combination), 단백질간 상호 작용, 단백질간 상호 작용 가능성

【명세서】

【발명의 명칭】

도메인 조합 정보를 이용한 단백질간의 상호 작용 예측 방법 및 예측 시스템{SYSTEM AND METHOD FOR PREDICTING THE INTERACTION BETWEEN PROTEINS BASED ON DOMAION COMBINATION}

【도면의 간단한 설명】

도 1은 종래의 도메인에 기반한 단백질 상호 작용 예측 모델을 도시한 도면.

도 2은 본 발명에 따른 도메인 조합 쌍의 예를 도시한 도면.

도 3은 본 발명에 따라 단백질간의 상호 작용을 예측하는 과정을 도시한 흐름도.

도 4는 본 발명에 따라 도메인 조합이 만들어질 때, 각 원소들이 어느 카테고리에 속하는지를 도시한 도면.

도 5는 상호 작용이 있는 것으로 알려진 집단과 상호 작용이 없다고 추정되는 집단을 대상으로 한 출현 빈도 배열 원소 값의 분포를 도시한 도면.

도 6은 본 발명에 따른 단백질의 상호 작용 예측 시스템을 도시한 도면.

도 7은 본 발명에 따른 단백질 상호 작용 예측 방법을 수행하는 데 채용될 수 있는 범용 컴퓨터 장치의 내부 블록도.

<도면의 주요 부분에 대한 부호의 설명>

600: 단백질의 상호 작용 예측 시스템

610: 확률 정보 저장부

620: 확률식 결정부

630: 확률식 연산부

【발명의 상세한 설명】

【발명의 목적】

【발명이 속하는 기술분야 및 그 분야의 종래기술】

<13> 본 발명은 단백질 상호 작용 예측 시스템 및 방법에 관한 것으로, 보다 자세하게는 도메인 조합 쌍에 기반하여 단백질의 상호 작용 확률을 예측하기 위한 새로운 확률적 단백질 상호 작용 예측 시스템 및 방법에 관한 것이다.

<14> 세포 내에서 일어나는 신호 전달, 세포 주기, 분화, DNA 복제 및 전사, 번역, 대사 등 거의 모든 반응들은 수많은 단백질의 상호작용을 통해 수행되고 조절 된다. 따라서, 단백질간의 상호 작용을 밝혀 세포 내에서의 생화학적 현상과 기전을 연구하는 것은 현대 생화학, 분자생물학의 주요한 연구 대상이 되고 있다.

<15> 종래 단백질간의 상호 작용을 연구함에 있어, 실험을 통하지 않고 계산적으로 예측하는 여러 가지 방법들이 연구되어 왔다. 일례로서, 가공하지 않은 단백질 서열로부터 직접 단백질-단백질 상호 작용에 영향을 끼치는 요인들을 발견하고 분석하는 것이 한 가지 접근 방법이며, 또 달리 단백질의 3차 구조나 물리화학적 특성을 분석함으로써 단백질 상호 작용을 예측하는 시도도 있었다.

<16> 도메인에 기반한 단백질-단백질 상호 작용 예측도 현재까지 활발히 연구되지 않았으나, 하나의 접근법이 될 수 있고 몇몇 시도가 이루어져왔다. 그 동안 도메인에 기반한 단백질-단백질 상호 작용 예측이 활발하지 않았던 것은 도메인이 복잡한 단백질-단백질 상호 작용의 모든 세부 사항들을 설명할 수는 없기 때문이며, 도메인에 대한 축적된 데이터가 아직 충분하지 않아 그 정확도 측면에서 신뢰성이 떨어지는 것도 한 가지 요인으로 볼 수 있다. 그러나, 오

늘날 인터넷 등의 통신망의 발달은 다량의 단백질 정보의 축적을 가져와 단백질 데이터에 기반한 단백질의 구조와 기능을 계산함에 있어 보다 효율성을 향상시키고 있다.

<17> 하지만, 종래의 도메인에 기반한 단백질 상호 작용 예측 모델들은 단백질-단백질 상호 작용이 독립적으로 발생하는 단일 도메인 쌍(single domain pair)의 결합에 의해 유발된다고 한정함으로써, 복수의 도메인들이 합동으로 단백질 상호 작용에 영향을 미친다는 점을 제대로 반영하지 못하였다. 왜냐하면 단백질의 상호 작용에는 단일 도메인 쌍뿐만 아니라, 동일 단백질 내에서 이에 연결되어 있는 다른 도메인들도 도메인 쌍의 형성에 영향을 미칠 수 있고 복수의 도메인들이 합동으로 단백질 상호 작용에 영향을 미칠 수 있기 때문이다. 그러므로, 도메인 쌍의 출현 빈도만을 추출하여 단백질의 상호 작용을 예측할 경우 단백질의 상호 작용에 함께 영향을 미치는 다른 도메인들의 변인을 간과하게 된다. 이러한 오류를 극복하기 위해서는 단일 도메인 쌍의 형성에 참여하지 않더라도 단백질을 구성하고 있는 다른 도메인들까지도 함께 고려하여 단백질의 상호 작용을 분석할 필요가 있다.

【발명이 이루고자 하는 기술적 과제】

<18> 종래의 도메인 기반 단백질 상호 작용 예측 모델이 가지는 제약성을 극복하기 위하여, 본 발명에서는 단백질 상호 작용은 복수의 도메인 쌍이나 도메인 조합(domain combination) 간의 상호 작용의 결과로 인식한다.

<19> 따라서, 본 발명의 목적은 단백질간 상호 작용에 영향을 미치는 다른 도메인들의 존재까지도 포괄적으로 분석에 고려하여 단백질간 상호 작용을 보다 정확하게 예측할 수 있도록 하는 것이다.

<20> 본 발명의 다른 목적은 종래의 기술이 주로 스코어링 시스템(scoring system)에 기반한 단순히 스코어(score) 값을 제공하는데 반해서, 상호 작용 가능성에 대한 확률 값을 제시함으로써 보다 현실적인 정보를 제공하는 것이다.

<21> 본 발명의 또 다른 목적은 적은 비용 및 시간으로 단백질 상호 작용 예측을 가능하게 하는 것이다.

<22> 본 발명의 또 다른 목적은 단백질간 상호 작용이 없을 것으로 가정된 임의의 단백질 집합(Random Protein Pair)에 대한 정보를 추가적으로 사용함으로써, 예측의 정확도를 높이는 것이다.

【발명의 구성 및 작용】

<23> 이하에서는 도메인 조합의 관점에서 단백질의 상호 작용을 예측하는 방법 및 시스템을 첨부된 도면을 참조하여 상세하게 설명한다.

<24> 도 1은 종래의 도메인에 기반한 단백질 상호 작용 예측 모델을 도시한 것이다. 도 1에서 보는 바와 같이, 단백질-단백질 상호 작용이 독립적으로 발생하는 단일 도메인 쌍(single domain pair)의 결합에 의해 유발된다고 한정한 종래의 방식은 단백질의 상호 작용이 독립적으로 발생하는 단일 도메인 쌍(Single Domain Pair)의 형성에 의해 유발된다는 가정에 기초한 것이다. 도 1에서는 각각 3개와 2개의 단일 도메인을 갖는 단백질이 상호 작용하는 경우, 잠재적인 상호 작용 도메인 쌍(Potentially Interacting Domain(PID) Pair)들을 보여주고 있다.

<25> 다음으로, 본 발명에서 제안하고 있는 예측 모델을 설명하기 전에 도메인 조합(Domain Combination)과 도메인 조합 쌍(Domain Combination Pair)의 개념을 먼저 설명한다. 편의상 도메인 조합은 간단히 *dc*, 도메인 조합 쌍은 *dc-pair*로 표기하기로 한다.

<26> 한 단백질 *p*가 복수의 도메인을 가지고 있다면, 도메인 조합은 단백질 *p*의 도메인 집합으로부터 만들어질 수 있는 모든 가능한 도메인 그룹이 된다. 여기서 그룹은 적어도 하나의 도메인을 반드시 포함하는 것으로 한다. 즉, 단백질 *p*의 모든 가능한 도메인 조합의 집합은 다음과 같이 정의된다.

<27>
$$dc(p) = Power(domain(p)) - \emptyset$$

【수학식 1】

<28> 여기에서, *domain*(*p*)는 단백질(protein) *p*의 도메인 집합을 나타낸다. [수학식 1]식에서 보듯이, 도메인 조합은 도메인 집합의 파워셋 P(A)에서 공집합(ø)을 뺀 값이다. 공집합 ø가 제거되므로 단백질이 *n* 개의 서로 다른 도메인을 가지고 있다면, 2^n-1개의 도메인 조합이 얻어진다.

<29> 본 발명에서 제시하는 예측 모델에서는 도메인 조합 쌍을 단백질 상호 작용의 기본 단위로 간주하며, 동일 단백질 안의 하나 이상의 복수의 도메인 조합 쌍이 연합하여 단백질 상호 작용에 영향을 주는 것으로 가정한다. 즉, 두 개의 단백질이 서로 상호 작용을 한다면, 두 단백질 간의 상호 작용은 각 단백질의 도메인 조합 간의 상호 작용의 결과로 해석한다. 이 관계를 설명하기 위하여, 두 단백질로 이루어진 도메인 조합 쌍 개념을 도입하였다. 두 단백질 *p*, *q*에서 모든 가능한 도메인 조합 쌍의 집합의 정의는 다음과 같다.

<30>
$$dc\text{-}pair(p, q)=\{\langle dc1, dc2\rangle \mid \langle dc1, dc2\rangle \in dc(p)\times dc(q) \text{ or } dc(q)\times dc(p)\}$$

【수학식 2】

<31> 즉, 두 단백질 *p*, *q*가 각각 *n*, *m*개의 다른 도메인을 가지고 있을 경우, 2^n-1, 2^m-1개의 서로 다른 도메인 조합 쌍(*dc-pairs*)를 얻게 된다.

<32> 도 2는 본 발명에 따른 도메인 조합 쌍의 예를 도시한 것이다. 도 2에서는 각각 3개와 2개의 도메인을 갖는 단백질이 상호 작용하는 경우, 잠재적인 상호 작용 도메인 조합 쌍 (Potentially Interacting Domain Combination(PIDC) Pair)들을 보여주고 있다. 도 1과 대비하여, 도 2는 기존 방법에서 사용했던 도메인 쌍 기반 방법과 도메인 조합 쌍 기반 접근 방법의 차이점을 보여준다. 향후 인터넷을 통한 상호 작용 단백질 쌍의 정보가 축적이 누적되면, 중요한 *dc-pairs*를 추출하는 것이 가능할 것으로 예상된다. 또한 *dc-pairs*의 역할의 정도를 정확히 결정하기 위해서는 적절한 가중치(weight) 부여가 매우 중요할 것으로 판단되며 이것에 관한 자세한 사항은 후술하기로 한다.

<33> 도 3은 본 발명에 따른 단백질 상호 작용 예측 방법에 관한 주요 흐름도를 도시한 것이다.

<34> 본 발명에서 제안된 예측 방법은 크게 예측을 준비하는 과정과 예측을 실제로 수행하는 과정으로 구성되어 있다.

<35> 예측 준비 과정은 다시 세 개의 단계를 포함한다.. 첫 번째 단계(S310)에서는 상호 작용이 있는 것으로 알려진 단백질쌍(이하, 상호 작용 단백질쌍) 집합(또는 모집단)과 상호 작용이 없는 것으로 추정되는 단백질쌍(이하, 비상호 작용 단백질쌍) 집합으로부터 각각 도메인 조합 정보와 그 출현 빈도를 추출한다. 이 정보들은 출현 확률 배열(Appearance Probability Matrix: AP matrix)라고 불리는 배열 구조에 저장된다.

<36> 두 번째 단계(S320)에서는 출현 확률 배열을 기반으로 단백질-단백질 상호 작용 예측 확률식을 정의한다. 이 확률식은 미 정의된 상수를 포함하게 되며 이 상수는 최대 가능도 추정(maximum likelihood estimation)을 적용하여 결정한다. 마지막 세 번째 단계(S330)에서는 상

호 작용이 있는 것으로 알려진 단백질 쌍 집합과 상호 작용이 없는 것으로 추정되는 단백질 도메인 쌍 집합의 확률 값 분포를 얻게 된다.

<37> 두 번째 과정에서는 첫 번째 과정에서 얻어진 분포에 기초하여, 단백질-단백질 상호 작용을 예측하는 또 다른 확률식이 정의되며, 이 확률식을 이용하여 확률을 계산한다. 이 확률은 단백질-단백질 상호 작용을 예측하는 최종 확률이다.

<38> 이하, 도 3을 참고로 하여, 각 단계들에서 수행되는 과정을 상세히 설명하기로 한다.

<39> 먼저, 단계(S310)에서는 도메인 조합의 출현 확률을 계산한다. 본 단계에서는 상호 작용 예측을 위한 확률식을 정의하기 위하여 필요한 도메인 조합의 출현 빈도를 상호 작용 단백질쌍의 집합과 비상호 작용 단백질쌍의 집합으로부터 추출한다. 도메인 조합의 출현 빈도(appearance frequency), 더욱 구체적으로는 모집단에서 특정 도메인 조합이 발생하는 빈도에 관한 데이터 수집을 위하여 출현 확률 배열 (Appearance Probability matrix; AP matrix)을 구성한다.

<40> 주어진 단백질 쌍 집합에서, *n*개의 서로 다른 단백질{*p1*, *p2*, ..., *pn*}이 있을 때, 단백질의 도메인 조합은, *dc*(*p1*), *dc*(*p2*), ..., *dc*(*pn*)이 되며 이 조합의 합집합은 *m*개의 서로 다른 도메인 조합 {*dc1, dc2, ..., dcm*}을 구성하게 되어, *m*-by-*m* AP 배열이 생성된다. 배열에서 원소 AP_{ij}는 주어진 단백질 쌍 집합에서 도메인 조합 $<dc_a, dc_b>$ 출현 확률을 대표한다.

<41> 출현 확률 배열을 만들기 위하여, 먼저 가중치 빈도(WF: Weighted Frequency) 배열을 먼저 생성한다. 이때 각 열과 줄은 도메인 조합을 나타내며, 배열의 각 원소는 *dc-pair*를 나타낸다. WF 배열에서는, 주어진 단백질 쌍의 집합에서의 도메인 조합 출현 빈도가 등록된다.

원소 WF*ab*는 *도메인 조합* <*a*, *b*>의 가중치 출현 빈도(weighted appearance frequency)를 가지게 되며, 다음 [수학식 3]에 의하여 계산된다.

<42>

$$WFab = \sum_{\substack{For\ all\ protein\ pairs\ pi,qj \\ s.t.\ <a,b>\in dc\text{-}pair(pi,qj)}} \frac{1}{|dc(p_i)| \times |dc(q_j)|}$$

【수학식 3】

<43>

즉, *dc-pair* <*a*, *b*>를 포함하는 모든 단백질 쌍 <*pi*, *qj*>에서 $\frac{1}{|dc(p_i)| \times |dc(q_j)|}$ 값을 계산하여 더함으로써, [수학식 3]의 최종 결과가 계산된다.

<44> [수학식 3]을 {<*A*,*B*>,<*A*,*C*>,<*B*,*C*>}로 주어진 단백질 쌍 집합의 각 단백질의 도메인이 dA = {*a1*,*a2*}, dB = {*b1*}, dC = {*a1*,*c1*}로 구성되어 있는 예에 적용하면, 도메인 조합 <{b1},{a2}>은 *dc-pair(A,B)*에서만 출현하므로 그 값은 1/(|*dc*(*B*)|· |*dc*(*A*)|)가 된다. 그리고 *dc*(*A*)={{*a1*},{*a2*},{*a1*,*a2*}}, *dc*(*B*)={{*b1*}}, |*dc*(*A*)| = 3, |*dc*(*B*)| = 1이므로, 최종적으로 1/(|*dc*(*B*)|· |*dc*(*A*)|) 의 값은 1/3이 된다. 이런 방법으로 WF 배열의 모든 요소들을 계산할 수 있다. WF 배열이 생성된 후에 AP 배열의 각 원소 값의 계산은 다음 [수학식 4]에 따른다.

<45>

$$AP_{ij} = \frac{WF_{ij}}{\sum_{i,j} WF_{ij}}$$

【수학식 4】

<46> 즉, 도메인 조합 배열의 모든 원소 값을 더한 값으로 원소 값을 나누어서 출현 확률 배열을 얻는다.

<47> 이와 같이 얻어진 배열의 각 원소 값은 특정 도메인 조합이 해당 공간에서

출현할 확률을 나타내게 된다. 상호 작용이 있는 것으로 알려진 단백질 쌍 집합과 상호 작용이 없는 것으로 추정되는 단백질 도메인 쌍 집합에 대하여 각각의 AP 배열을 얻을 수 있다. 두 배열의 상당한 부분이 서로 겹쳐지지만, 형태가 꼭 일치할 필요는 없다.

<48> 그리고, 상호 작용이 있는 것으로 알려진 단백질 쌍 그룹과 상호 작용이 없는 것으로 추정되는 그룹 각각에 대해서 출현 확률을 구할 수 있으므로 얻어진 출현 확률 배열을 AP^i, AP^r 배열로 표시하고 이들의 공통 부분 $AP^i \cap AP^r$은 AP^c 배열로 나타낸다. 각 배열에 대한 좀 더 엄밀한 정의는 다음과 같다.

<49> AP^r: 상호 작용이 없는 것으로 추정되는 단백질 쌍 집합으로부터 얻어지는 AP 배열

<50> AP^i : 상호 작용이 있는 단백질 쌍 집합으로부터 얻어지는 AP 배열

<51> AP^c : $AP^i \cap AP^r$

<52> 일단, 상호 작용이 있는 것으로 알려진 쌍과 없는 것으로 추정되는 AP 배열이 얻어지면 *dc-pair*를 각각 그들이 속하는 그룹으로 분류할 수 있으며, AP^i, AP^r, AP^c 개념을 이용하여 각 범주(category)를 명명할 수 있게 된다. AP^i 배열 상에 나타나는 모든 dc-pair는 AP^i *dc-pair* 공간을 구성한다. 같은 방법으로, AP^r *dc-pair* 공간, AP^c *dc-pair* 공간이 정의된다.

<53> 다음으로 단계(S320)에서는 첫 과정에서 얻어진 두 개의 출현 확률 배열을 기반으로, 상호 작용을 모르는 단백질 쌍 <A,B>에 대한 확률을 예측하는 확률식이 정의되며, 이 확률식에 포함되는 미지의 상수가 결정된다.

<54> 먼저, 단백질 쌍 <A,B>로부터 [수학식 2]를 이용하여, 이들의 도메인 조합 *dc-pairs*를 산출한다. 많은 *dc-pair*들이 만들어지며, *dc-pairs* 공간에는 여러 개의 범주가 있으므로, 그 범주에 따라 다음과 같이 *dc-pairs*를 분류한다.

<55>

DC*c*(A,B)={*dc-pair* | *dc-pair*∈*dc-pair*(A,B) and appears in AP^c dc-pair space}

DC*r-c*(A,B)={*dc-pair* | *dc-pair*∈*dc-pair*(A,B) and appears in {AP^r– AP^c} space}

DC*i-c*(A,B)={*dc-pair* | *dc-pair*∈*dc-pair*(A,B) and appears in {AP^i– AP^c} space}

<56> 도 4는 AP^i, AP^r 공간에서 *dc-pair*(A, B)가 만들어질 때, 각 원소들이 어느 카테고리에 속하는지를 보여준다. 도 4에서 위 *dc-pair*(*A, B*)의 각 원소들은 특수 기호(*,△,×)로 표시된다.

<57> AP^c *dc-pair* 공간에서 발견되는 *DCc*(*A, B*) 도메인 조합을 대상으로 상호 작용 확률식을 아래의 [수학식 5]와 같이 정의할 수 있다.

<58> 이 확률은 *DCc*(*A, B*)가 AP^c *dc-pair* 공간에서 발견될 때 단백질 쌍 <A,B>가 서로 상호 작용할 확률을 의미한다. 상호 작용이 일어나는 사건과 일어나지 않는 사건을 표현하기 위하여 확률 변수 X를 도입하였다. [수학식 5]에서 1 값은 상호 작용이 일어나는 사건, 0 값은 상호 작용이 없는 사건을 나타낸다.

<59>

$$P(X=1|DCc(A,B)) = \frac{P(X=1)P(DC_c(A,B)|X=1)}{P(X=1)P(DC_c(A,B)|X=1)+P(X=0)P(DC_c(A,B)|X=0)}$$

【수학식 5】

<60> 그리고, $P(X=1)$, $P(X=0)$, $P(DCc(A,B)|X=1)$, $P(DCc(A,B)|X=0)$의 정의는 다음과 같다.

<61>

$$P(X=1) = \frac{k \cdot I_{total} \cdot \sum_{i,j} (AP_I^c)_{ij}}{k \cdot I_{total} \cdot \sum_{i,j} (AP_I^c)_{ij} + (1-k) \cdot R_{total} \cdot \sum_{i,j} (AP_R^c)_{ij}}$$

$$P(X=0) = \frac{(1-k) \cdot R_{total} \cdot \sum_{i,j} (AP_R^c)_{ij}}{k \cdot I_{total} \cdot \sum_{i,j} (AP_I^c)_{ij} + (1-k) \cdot R_{total} \cdot \sum_{i,j} (AP_R^c)_{ij}}$$

$$P(DCc(A,B)|X=1) = |DCc(A,B)|\ ! \prod_{(i,j)|(i,j)\in DCc(A,B))} \frac{(AP_I^c)_{ij}}{\sum_{i,j} (AP_I^c)_{ij}}$$

$$P(DCc(A,B)|X=0) = |DCc(A,B)|! \prod_{(i,j)|(i,j)\in DCc(A,B))} \frac{(AP_R^c)_{ij}}{\sum_{i,j} (AP_R^c)_{ij}}$$

<62> 이 때, $P(X=1)$는 AP^c에 존재하는 총 *dc-pair* 공간에서 상호 작용이 있는 단백질 쌍으로부터 만들어진 *dc-pair* 공간을 나타내며, $P(X=0)$은 AP^c의 도메인 조합 공간에서 상호 작용이 없다고 추정되는 단백질 쌍으로부터 생성된 *dc-pair* 공간을 나타낸다. I_{total}과 R_{total}은 상호 작용이 있는 단백질 쌍과 상호 작용이 없는 것으로 간주되고 있는 단백질 쌍의 총 개수를 각각 나타낸다. 식에서 상수 k는 자연계에서 I_{total}과 R_{total} 의 비율을 나타내며 이 값을 정확하게 알 수 없으므로, 최대 가능도 추정(maximum likelihood estimation) 적용을 통하여 결정한다.

<63> $P(DCc(A,B)|X=1)$는 AP^i 공간에서 DCc(A, B)에 속하는 *dc-pairs* 집합이 만들어질 확률이고, $P(DCc(A,B)|X=0)$는 AP^r공간에서 DCc(A, B)에 속하는 *dc-pairs* 집합이 만들어질 확률이다. $AP_I{}^c$와 $AP_R{}^c$ 는 각각 상호 작용이 있는 *dc-pair* 공간과 상호 작용이 없는 것으로 간주되고 있는 *dc-pair* 공간에서 AP^i를 의미한다. 동일하게, DCi-c(A,B) 도메인 조합을 대상으로 얻어질 확률식은 다음과 같다.

<64>

$$P(X=1|DCi\text{-}c(A,B)) = \frac{P(X=1)P(DCi\text{-}c(A,B)|X=1)}{P(X=1)P(DCi\text{-}c(A,B)|X=1) + P(X=0)P(DCi\text{-}c(A,B)|X=0)}$$

【수학식 6】

<65> $P(X=1/DCi\text{-}c(A,B))$ =

<66>
$$\frac{P(X=1)P(DC_i\text{-}c(A,B)|X=1)}{P(X=1)P(DC_i\text{-}c(A,B)|X=1)+P(X=0)P(DC_i\text{-}c(A,B)|X=0)}$$

<67> 위 식에서, $P(X=1)$, $P(X=0)$는 각각 1, 0이 되어 최종적으로 얻어지는 확률은 1이 된다. [수학식 5] 및 [수학식 6]을 이용하여, DC$c(A,B)$ *dc-pairs*를 갖는 (A, B) 단백질 쌍의 상호 작용 가능성 확률(Primary Interaction Probability; PIP)은 다음 식에 의하여 계산된다.

<68>
$$PIP(A,B) = 1 - \frac{AP^c}{AP^r}(1-P(X=1|DCc(A,B))$$

【수학식 7】

<69> 단계(S320)에서 PIP 최종식이 얻어지면, [수학식7]에 따라 상호 작용이 있는 단백질 쌍과 없는 것으로 간주된 쌍 집합에 대한 PIP 값을 계산할 수 있다(단계 S330 및 단계 S340).

<70> 본 발명에 따른 또다른 실시예로서 그들을 비교하기 위하여 분포를 정규화할 수 있다. 한편 PIP 함수는 단백질 쌍을 실수 0 내지 1 범위 안에 전사시키는 함수의 일종으로 해석할 수 있다.

<71> 일단 단백질 상호 작용 예측 분포가 얻어지면, 이 분포에 대한 2-카테고리 분류(two-category classification) 기법 적용이 가능하다. 즉, 임의로 주어진 단백질 쌍에 대하여, 그들이 상호 작용을 할지 안 할지 예측하기 위해서는 그 단백질 쌍의 PIP 값이 어느 분포에 속할지를 결정해야 한다. 2-카테고리 분류(two-category classification)의 많은 기법이 있지만, 이를 확률적으로 표현하기 위하여, 단백질 쌍의 조건부 확률을 계산하여 어떤 카테고리에 속하는지를 결정하였다.

<72> 일례로서, <A,B> 단백질 쌍의 상호 작용 확률을 예측하기 위하여 먼저 DC(A,B)를 계산하고, [수학식 7]을 이용하여 PIP(A,B)를 계산한다. 그리고 최종 <A, B> 단백질 쌍의 상호 작용 다음 확률식(수학식 8)에 의하여 계산된다.

<73>

【수학식 8】

$$P(X=1|p=PIP(A,B)) = \frac{P(X=1)\,P(p=PIP(A,B)|X=1)}{P(X=1)\,P(p=PIP(A,B)|X=1)+P(X=0)\,P(p=PIP(A,B)|X=0)}$$

<74> 여기에서, $P(X=1)$, $P(X=0)$, $P(p=PIP(A,B)|X=1)$, $P(p=PIP(A,B)|X=0)$는 각각

<75>

$$P(X=1) = \frac{k\cdot\sum_{i=1}^{m} freq_i^x}{k\cdot\sum_{i=1}^{m} freq_i^x + (1-k)\sum_{i=1}^{m} freq_i^y}$$

$$P(X=0) = \frac{(1-k)\cdot\sum_{i=1}^{m} freq_i^y}{k\cdot\sum_{i=1}^{m} freq_i^x + (1-k)\sum_{i=1}^{m} freq_i^y}$$

$$P(p=PIP(A,B)|X=1) = \frac{freq^x_{PIP(A,B)}}{\sum_{i=1}^{m} freq_i^x}$$

$$P(p=PIP(A,B)|X=0) = \frac{freq^y_{PIP(A,B)}}{\sum_{i=1}^{n} freq_i^y}$$

<76> 으로 정의된다.

<77> $P(X=1)$는 총 단백질 쌍 수 중 상호 작용이 있다고 알려진 쌍의 비율을 나타내며, $P(X=0)$는 총 단백질 쌍 수 중에 상호 작용이 없다고 추정되는 쌍의 비율을 나타낸다. 또한, $freq_i^x$와 $freq_i^y$는 각각 PIP_i^x를 가지는 샘플이 상호 작용이 알려진 집합에 출현하는 빈도와 PIP_i^y를 가지는 샘플이 상호 작용이 없다고 추정되는 집합에 출현하는 빈도이다. 또한, 상수 k는 [수학식 5]에서 쓰여지는 것과 동일하다.

<78> *P(p=PIP(A,B)/X=1)*는 상호 작용이 있는 모 집단 내에서 확률변수 *p* 값이 PIP(A,B)가 될 확률을 의미한다. 마찬가지로, *P(p=PIP(A,B)/X=0)*는 상호 작용이 없다고 추정되는 모 집단 내에서 확률변수 *p* 값이 PIP(A,B)가 될 확률을 의미한다. 경우에 따라서 PIP(A,B)값을 갖는 샘플이 모집단 내에 존재하지 않는 수도 있어, 이 경우에는 사전에 정의된 범위 안에 존재하는 *p* 값을 대신 사용한다.

<79> 이어서, 본 발명에 따라 제안된 예측 방법을 위한 검증 단계들이 설명된다. 검증을 위하여, 다음과 같은 2개의 단백질 쌍 데이터를 준비하였다. 상호 작용이 알려진 단백질 쌍 집합은 DIP 데이터베이스 (http://dip.doe-mbi.ucla.edu) 의 효모(yeast)에서 총 15,174개의 상호 작용이 보고된 단백질 쌍 (효모(yeast)20030202.1st)을 준비하였다. 반면에, 상호 작용이 없다고 추정되는 단백질 쌍은 도메인 정보가 알려진 단백질 쌍 집단에서, 상호 작용이 알려진 단백질 쌍 집단을 제거하는 방식으로, 임의로 생성되었다. 총 단백질 각각에 대한 도메인의 정보는 PAD(http://www.ebi.ac.uk/proteome/)(R. Apweiler, M. Biswas, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, E. V. Kriventseva, V. Mittard, N. Mulder, I. Phan and E. Zdobnov, Proteome Analysis Database: online application of Interpro and CluSTr for the functional classification of proteins in whole genomes. *Nucleic Acids Res.*, 29(1):44-48, 2001. 참조)에서 추출(4932.SPC)하였다. 입증의 편의를 위하여, 상호 작용이 없는 것으로 추정된 단백질 쌍의 경우에는 상호 작용이 보고된 단백질 쌍과 같은 수의 단백질 쌍을 준비하였다. 그럼에도 불구하고, 추정되는 집단 안에 상호 작용이 있는 단백질 쌍이 완전히 제거된 상태는 아니다. 하지만 만일 전체 단백질 쌍 공간 안에 상호 작용 하는 단백질 쌍이 아주 드물다고 추측한다면, 본 예측 모델에서 사용된 상호 작용이 없다고 추정되는 집단으

로도 충분할 것이며, 입증 결과가, 이러한 방법으로 상호 작용이 없다고 추정되는 집단을 생성하고 사용하는 것이 적절하다는 것을 보일 것으로 예상된다.

<80> 이상의 방법으로 2개의 집단을 준비한 후에는, 각각을 학습 집단과 검증 집단으로 나누었다. 학습 집단으로 상호 작용이 있는 것으로 알려진 전체 단백질 쌍의 80%를 사용했을 때, n×n 크기의 AP^i와 m×m 크기의 AP^r이 생성되었다. 도 5는 상호 작용이 있는 것으로 알려진 집단과 상호 작용이 없다고 추정되는 집단을 대상으로 한 AP 배열 원소 값의 분포를 보여주고 있다. 본 발명에 따른 일례에서는, AP^i와 AP^r의 크기를 각각 13,000×3,000으로 하여 검증을 수행하였다.

<81> 각 배열의 크기는 매우 방대하고, 배열의 각 요소들은 출현 확률을 나타내기 때문에, 각 요소의 값은 보통 매우 작은 값들이다. 다음 단계에서 PIP값을 계산할 때 정확한 결과를 얻기 위하여, [수학식 5]의 계산순서에 따라 약간의 변형도 가능하다.

<82> AP^i, AP^r 배열이 얻어진 후에는 [수학식 7]을 적용하여 PIP 값의 분포를 얻을 수 있다. 도 5에서 상호 작용이 있는 것으로 보고된 단백질 쌍 집합의 분포(A)와 상호 작용이 없는 것으로 추정되는 단백질 쌍 집합의 분포(B)를 나타낸다. 각 집단의 PIP 값은 0 내지 1 사이에 중복되어 위치한다. 그러나 상호 작용이 보고된 집단의 PIP 값들은 대부분 1 가까이 있으며 상호 작용이 없다고 추정되는 집단의 PIP 값은 0 주위에 위치한다. 이것은 PIP 값이 상호 작용이 보고된 집단과 상호 작용이 없다고 추정되는 집단을 나누는 좋은 분류자(classifier)가 됨을 나타낸다. PIP 값의 분포를 다양한 2-카테고리 분류(2-category classification)방식을 적용하여 분류할 수 있다. 여기에서는 본 예측 모델의 유효성을 검사하기 위하여, [수학식 9] 값을 최소화하는 새로운 hybrid classification 방식을 고안하여 분류한 후 예측에 사용하였다.

<83>

$$P(e) = \sum_{\{i,j \mid PIP_i^x = PIP_j^y\}} Min[p_i^x, p_j^y]$$

$$P_i^x = \frac{freq_i^x}{\sum_{i=1}^{m} freq_i^x}$$

$$P_i^y = \frac{freq_i^y}{\sum_{i=1}^{n} freq_i^y}$$

【수학식 9】

<84> 따라서, 에러 확률 $P(e)$는 두 집단 간의 PIP 값이 중복되는 경우가 적을수록 감소한다. 본 모델의 유효성을 테스트하기 위하여 베이즈 규칙(Bayes rule)을 사용하여 민감도와 특이도를 측정하였다. 상호 작용이 알려진 단백질 쌍과 없다고 추정되는 단백질 쌍의 실험세트(learning set)을 이용하여, 3번 반복으로 테스트하여 보았다. 상호 작용이 알려진 단백질 쌍 전체 수 중 80%의 단백질 쌍을 실험세트로 사용하였을 때 약 90%의 민감도(sensitivity)와 약 80%의 특이성(specificity)이 얻어져 기존의 방식에 비하여 현저하게 예측의 정확도가 개선되는 것이 확인되었다. 여기서 민감도라 함은 전체 테스트 샘플에서 상호 작용이 있는 것에 대해서 상호 작용이 있는 것으로 예측하는 비율을 의미하고 특이성이라 함은 전체 테스트 샘플에서 상호 작용이 없는 것에 대해서 상호 작용이 없는 것으로 예측하는 비율을 의미하는 것으로 이 값이 높을수록 예측의 정확도가 좋음을 의미한다.

<85> 현재 보고된 상호 작용이 있다고 보고된 단백질 쌍 집단과 본 발명에서 임의로 생성된 상호 작용이 없다고 추정되는 단백질 쌍 안에는 실험적인 에러를 포함할 수 있으므로, 본 발명에서 계산된 민감도와 특이도에도 에러가 있을 수 있다. 얼마나 많은 데이터가 오류인지는 단정하기 어렵지만, 본 발명의 테스트 결과로 볼 때, 에러 데이터는 많은 부분을 차지하지 않는 것으로 추정되며, 본 예측 모델이 유효하다고 결론지을 수 있다. 이러한 결과는 상호 작용의

기본 단위로 *dc-pair*를 채택한 점과 분류를 위하여 PIP식을 사용한 것이 주효한 것으로 판단된다.

<86> 도 6은 본 발명에 따른 단백질의 상호 작용 예측 시스템(600)을 도시한 것이다. 도 6에서 보는 바와 같이, 예측 시스템(600)은 확률 정보 저장부(610), 확률식 결정부(620), 확률식 연산부(630)를 포함한다.

<87> 확률 정보 저장부(610)는 상호 작용 단백질쌍 모집단과 비상호 작용 단백질쌍 모집단의 각 집단으로부터 각각 선정한 특정 도메인 조합의 출현 확률 정보를 추출하고 저장한다. 도메인 조합의 출현 확률 정보는 출현 확률 배열로 정의되고, 상기 출현 확률 배열의 요소 AP_{ij}는 [수학식 3] 및 [수학식 4]에 의해서 결정될 수 있으며, 이에 대한 보다 상세한 설명은 앞서의 실시예들을 참조하기 바란다.

<88> 확률식 결정부(620)는 저장된 도메인 조합의 출현 확률 정보를 이용하여 임의의 두 단백질이 상호 작용할 확률식을 결정하게 된다. 확률식에 관한 상세한 설명은 앞서의 실시예들에서 이미 상세히 설명된 바 있다.

<89> 확률식 연산부(630)는 결정된 확률식으로부터 임의의 두 단백질이 상호 작용할 확률을 구하게 된다. 확률식 연산부(630)는 상기 결정된 상호 작용 확률식을 상호 작용 단백질쌍 및 비상호 작용 단백질쌍의 양 모집단에 적용하여 확률값 및 그 확률값의 분포를 구하고, 상기 결정된 상호 작용 확률식을 임의의 주어진 단백질쌍에 적용하여 확률값을 구하고, 상기 모집단의 분포를 바탕으로 임의의 주어진 단백질쌍이 어느 모집단에 속할지의 확률을 계산한다.

<90> 본 발명의 실시예들은 다양한 컴퓨터로 구현되는 동작을 수행하기 위한 프로그램 명령을 포함하는 컴퓨터 판독 가능 매체를 포함한다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령,

데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 상기 매체는 프로그램 명령, 데이터 구조 등을 지정하는 신호를 전송하는 반송파를 포함하는 광 또는 금속선, 도파관 등의 전송 매체일 수도 있다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

<91> 도 7은 본 발명에 따른 단백질 상호 작용 예측 방법을 수행하는 데 채용될 수 있는 범용 컴퓨터 장치의 내부 블록도이다.

<92> 컴퓨터 장치(700)는 램(RAM: Random Access Memory)(702)과 롬(ROM: Read Only Memory)(703)을 포함하는 주기억장치와 연결되는 하나 이상의 프로세서(701)를 포함한다. 프로세서(710)는 중앙처리장치(CPU)로 불리기도 한다. 본 기술분야에서 널리 알려져 있는 바와 같이, 롬(730)은 데이터(data)와 명령(instruction)을 단방향성으로 CPU에 전달하는 역할을 하며, 램(702)은 통상적으로 데이터와 명령을 양방향성으로 전달하는 데 사용된다. 램(702) 및 롬(703)은 컴퓨터 판독 가능 매체의 어떠한 적절한 형태를 포함할 수 있다. 대용량 기억장치(Mass Storage)(704)는 양방향성으로 프로세서(701)와 연결되어 추가적인 데이터 저장 능력을 제공하며, 상기된 컴퓨터 판독 가능 기록 매체 중 어떠한 것일 수 있다. 대용량 기억장치

(704)는 프로그램, 데이터 등을 저장하는데 사용되며, 통상적으로 주기억장치보다 속도가 느린 하드디스크와 같은 보조기억장치이다. CD 롬(706)과 같은 특정 대용량 기억장치가 사용될 수도 있다. 프로세서(701)는 비디오 모니터, 트랙볼, 마우스, 키보드, 마이크로폰, 터치스크린형 디스플레이, 카드 판독기, 자기 또는 종이 테이프 판독기, 음성 또는 필기 인식기, 조이스틱, 또는 기타 공지된 컴퓨터 입출력장치와 같은 하나 이상의 입출력 인터페이스(705)와 연결된다. 마지막으로, 프로세서(701)는 네트워크 인터페이스(707)를 통하여 유선 또는 무선 통신 네트워크에 연결될 수 있다. 이러한 네트워크 연결을 통하여 상기된 방법의 절차를 수행할 수 있다. 상기된 장치 및 도구는 컴퓨터 하드웨어 및 소프트웨어 기술 분야의 당업자에게 잘 알려져 있다.

<93> 상기된 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있다.

<94> 지금까지 본 발명에 따른 몇몇 실시예에 관하여 설명하였으나, 본 발명의 범위에서 벗어나지 않는 한도 내에서는 여러 가지 변형이 가능함은 물론이다.

<95> 본 발명의 범위는 설명된 실시예에 국한되어 정해져서는 안되며, 후술하는 특허청구의 범위뿐 아니라 이 특허청구의 범위와 균등한 것들에 의해 정해져야 한다.

【발명의 효과】

<96> 본 발명에 따르면, 많은 비용과 시간이 소요되는 단백질 상호 작용 실험을 통하지 않고도 적은 비용 및 시간으로 대량의 상호 작용 예측이 가능하다는 효과를 얻을 수 있다.

<97> 본 발명에 따른 계산적 방법에 의한 단백질 상호 예측은 단시간 내에 예측된 정보를 이용하여 수많은 후보 단백질 중에 우선 순위를 부여하여 실험 착수를 진행할 수 있다는 효과를 얻을 수 있다.

<98> 본 발명에 따르면, 단시간 내에 대규모 단백질 쌍에 대해서 상호 작용 가능성을 예측할 수 있어 이를 기반으로 대규모 단백질 상호 작용 네트워크 구성이 용이하고 다시 이를 기반으로 수많은 단백질 중에서 중요한 단백질을 추정하고 검증하는 데 활용할 수 있다.

<99> 본 발명에 따르면, 미지의 단백질에 대한 기능을 추정하는 것과 같은 단백질 동정 시에 기본적인 계산적 접근방법을 제공할 수 있다.

<100> 본 발명에 따른 예측 틀은 생물학자들이 그들의 연구 분야에서 유사한 경우를 만났을 때 참고 모델로 이용될 것이다.

<101> 본 발명에 따르면, 단백질간 상호 작용에 영향을 미치는 다른 도메인들의 존재까지도 포괄적으로 분석에 고려하므로 단백질간 상호 작용을 보다 정확하게 예측할 수 있다.

<102> 본 발명에 따르면, 단백질간 상호 작용이 없을 것으로 가정된 임의의 단백질 집합(Random Protein Pair)에 대한 정보를 추가적으로 사용하여 확률식을 정의한 후 다시 이들 정보를 피드백 적용하여 확률식을 자체적으로 검정할 수 있도록 하여, 예측의 정확도를 높였다.

<103> 본 발명에 따르면, 임의로 주어지는 단백질 쌍이 상호 작용 단백질 쌍 모집단에 속할지의 가능성을 확률값으로 예측받을 수 받을 수 있으므로, 단백질의 상호 작용에 대한 현실적인 예측을 기대할 수 있다.

【특허청구범위】

【청구항 1】

상호 작용 단백질쌍 모집단과 비상호 작용 단백질쌍 모집단의 각 집단으로부터 각각 선정한 특정 도메인 조합의 출현 확률 정보를 추출하고 저장하는 단계;

상기 저장된 도메인 조합의 출현 확률 정보를 이용하여 임의의 두 단백질이 상호 작용할 확률식을 결정하는 단계; 및

상기 결정된 확률식으로부터 임의의 두 단백질이 상호 작용할 확률을 구하는 단계

를 포함하는 것을 특징으로 하는 단백질의 상호 작용 예측방법.

【청구항 2】

제1항에 있어서,

특정 도메인 조합의 출현 확률 정보를 추출하고 저장하는 상기 단계는

가중치 빈도 배열을 생성하는 단계; 및

상기 가중치 빈도 배열을 기초로 출현 확률 배열을 생성하는 단계

를 포함하는 것을 특징으로 하는 단백질의 상호 작용 예측방법.

【청구항 3】

제1항에 있어서, 상기 도메인 조합의 출현 확률 정보는 출현 확률 배열로 정의되고, 상기 출현 확률 배열의 요소 AP_{ij}는

$$AP_{ij} = \frac{WF_{ij}}{\sum_{i,j} WF_{ij}}$$

$$WF_{ab} = \sum_{\substack{\text{For all protein pairs } p_i, q_j \\ \text{s.t.} <a,b> \in dc\text{-}pair(p_i, q_j)}} \frac{1}{|dc(p_i)| \times |dc(q_j)|}$$

로 결정되는 것을 특징으로 하는 단백질의 상호 작용 예측 방법.

【청구항 4】

제1항에 있어서, 임의의 두 단백질이 상호 작용할 확률을 구하는 상기 단계는

상기 결정된 상호 작용 확률식을 상기 상호 작용 단백질쌍 모집단 및 상기 비상호 작용 단백질쌍 모집단에 적용하여 확률값 및 상기 확률값의 분포를 구하는 단계;

상기 결정된 상호 작용 확률식을 임의의 주어진 단백질쌍에 적용하여 확률값을 구하는 단계; 및

모집단의 분포를 바탕으로 임의의 주어진 단백질쌍이 어느 모집단에 속할지의 확률을 계산하는 단계

를 포함하는 것을 특징으로 하는 단백질의 상호 작용 예측 방법.

【청구항 5】

상호 작용 단백질쌍 모집단과 비상호 작용 단백질쌍 모집단의 각 집단으로부터 각각 선정한 특정 도메인 조합의 출현 확률 정보를 추출하고 저장하는 확률 정보 저장부;

상기 저장된 도메인 조합의 출현 확률 정보를 이용하여 임의의 두 단백질이 상호 작용할 확률식을 결정하는 확률식 결정부; 및

상기 결정된 확률식으로부터 임의의 두 단백질이 상호 작용할 확률을 구하는 확률식 연산부

를 포함하는 것을 특징으로 하는 단백질의 상호 작용 예측 시스템.

【청구항 6】

제4항에 있어서, 상기 확률식 연산부는

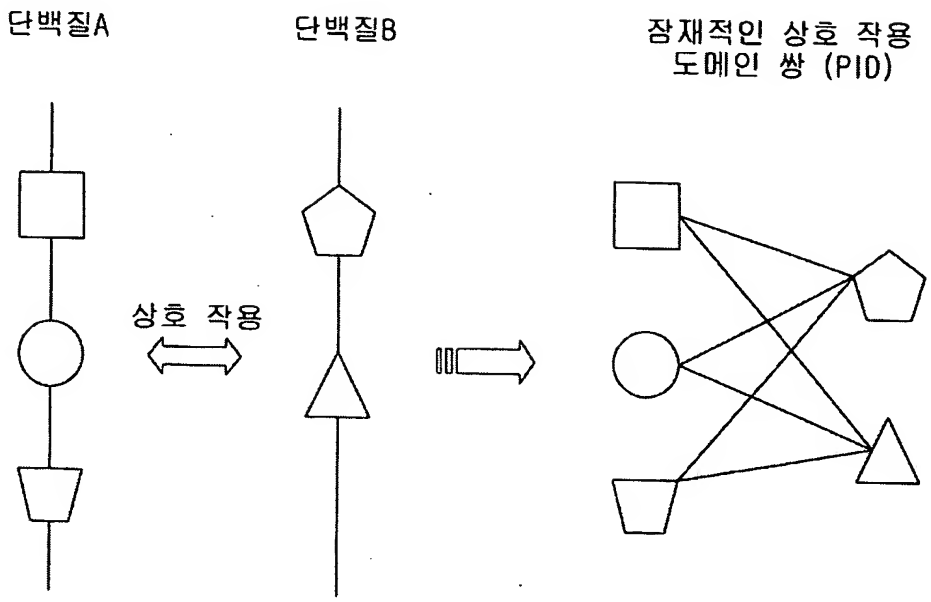
상기 결정된 상호 작용 확률식을 상기 상호 작용 단백질쌍 및 상기 비상호 작용 단백질쌍의 양 모집단에 적용하여 확률값 및 그 확률값의 분포를 구하고, 상기 결정된 상호 작용 확률식을 임의의 주어진 단백질쌍에 적용하여 확률값을 구하고, 상기 모집단의 분포를 바탕으로 임의의 주어진 단백질쌍이 어느 모집단에 속할지의 확률을 계산하는 것을 특징으로 하는 단백질의 상호 작용 예측 시스템.

【청구항 7】

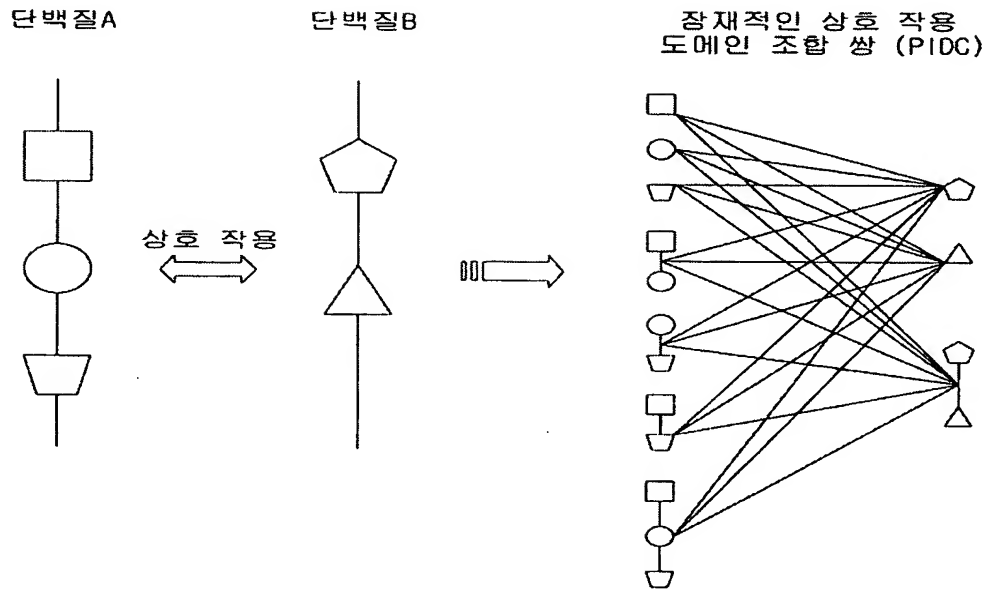
제1항 내지 제4항의 방법 중 어느 하나의 항에 따른 방법을 컴퓨터에서 구현하는 프로그램을 기록한 컴퓨터 판독 가능한 기록 매체.

【도면】

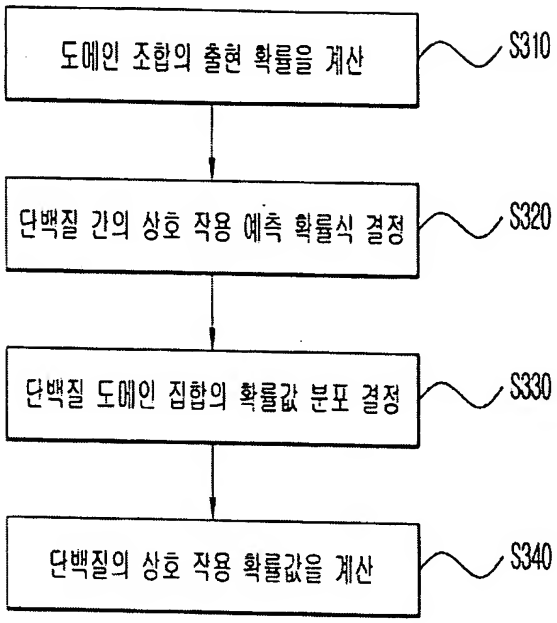
【도 1】


단백질A
단백질B
잠재적인 상호 작용
도메인 쌍 (PID)
상호 작용

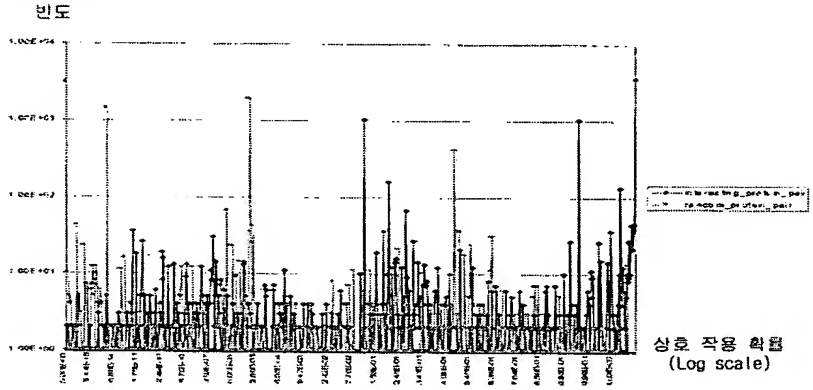
【도 2】


단백질A
단백질B
잠재적인 상호 작용
도메인 조합 쌍 (PIDC)
상호 작용

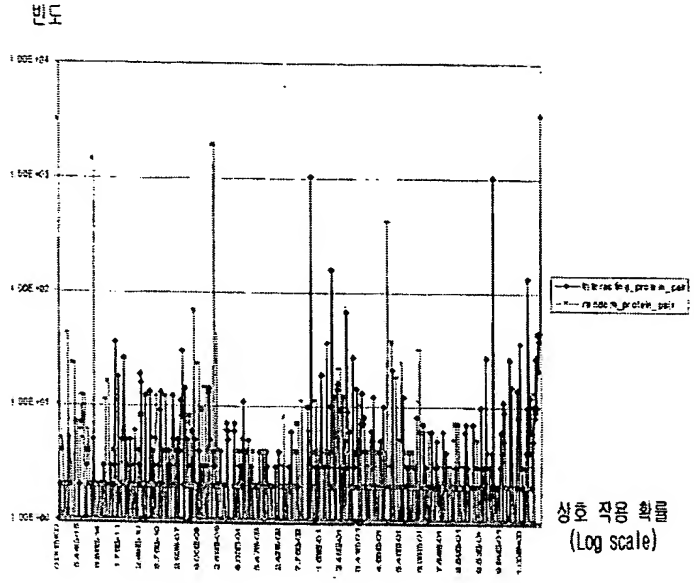
【도 3】


도메인 조합의 출현 확률을 계산
S310
단백질 간의 상호 작용 예측 확률식 결정
S320
단백질 도메인 집합의 확률값 분포 결정
S330
단백질의 상호 작용 확률값을 계산
S340

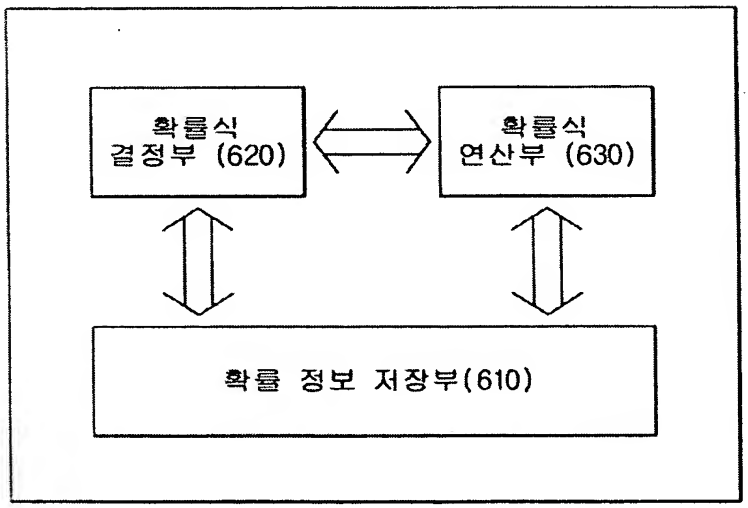
【도 4】


빈도
상호 작용 확률
(Log scale)

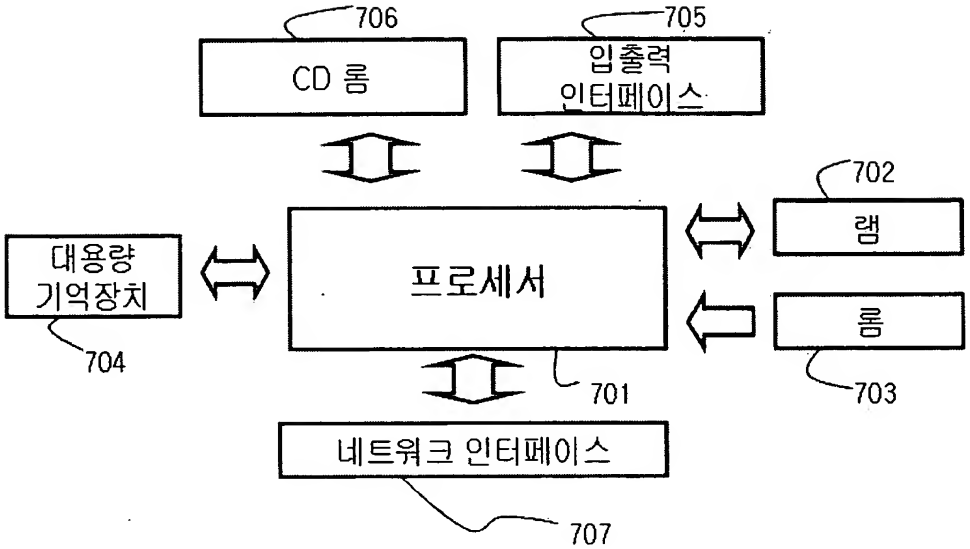
【도 5】



【도 6】



【도 7】


706
CD 롬
705
입출력
인터페이스
702
램
대용량
기억장치
704
프로세서
롬
703
701
네트워크 인터페이스
707
700